A rank graduation index to prioritise cyber risks

Un indice di graduazione per assegnare livelli di priorità ai rischi informatici

Paolo Giudici and Emanuela Raffinetti

Abstract In this paper we introduce a new methodology for estimating the risks of cyber attacks. In order to deal with the ordinal nature of the cyber risk response variable, an extension of linear regression models is proposed, by means of the rank tools. We also suggest a specific model evaluation measure, called *RG* (Rank Graduation), aiming at detecting the factors which mainly affect cyber risks. Finally, to shed light on the effectiveness of our proposal, we use our proposed methodology to rank real cyber loss data.

Abstract *In questo articolo, introduciamo una nuova metodologia per la stima dei rischi legati agli attacchi informatici. Allo scopo di superare le problematiche associate alla natura ordinale della variabile risposta, identificabile con il rischio informatico, proponiamo un'estensione dei modelli di regressione lineare basata sull'utilizzo dei ranghi. Infine, con l'obiettivo di individuare i fattori che principalmente incidono sul rischio informatico, una nuova misura di valutazione del modello, chiamata RG (Rank Graduation), viene presa in considerazione. L'articolo si conclude con un'interessante applicazione della metodologia proposta ai dati reali, che mette ulteriormente in evidenza la sua efficacia nel processo di classificazione dei rischi informatici.*

Key words: cyber risk, ordinal variables, rank-based methods

Paolo Giudici
Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia (Italy), e-mail: paolo.giudici@unipv.it

Emanuela Raffinetti
Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, 20122 Milan (Italy), e-mail: emanuela.raffinetti@unimi.it

1 Introduction

In the last few years the number of cyber attacks has surged, with a growth of about 30% between 2014 and 2017. The trend in 2018 follows a similar behavior, with 730 cyber attacks observed only in the first half of the year [2]. Cyber risks can be defined as "any risk emerging from the use of information and communication technology (ICT) that compromises the confidentiality, availability, or the integrity of data or services" (see e.g. [4]).

Financial institutions are encouraged by regulators to use statistical approaches to estimate the capital charge covering operational risk, which include cyber risks. This requires the presence of historical loss data, in a quantitative format. We remark that cyber events are typically expressed on ordinal scales. While the literature on the quantitative measurement of operational risks (see e.g. [3]), based on loss data, constitute a reasonably large body, that on cyber risk measurement and, especially, on ordinal cyber risk measurement, is very limited. Our contribution tries to fill this gap in the literature, providing a cyber risk model based on ordinal data. Specifically, given the ordinal nature of the target variable measuring the severity degree, a new approach that extends linear regression models is introduced. Furthermore, since an essential part of the cybersecurity management is to detect the main factors affecting the severity degree, it seems appropriate to validate the different models used for detecting the variables impacting on it through specific predictive accuracy measures.

Typically, the choice of the most suitable validation metric is strictly related to the the nature of the response variable to be predicted. Recently, a measure that is objective and not endogenous to the system itself was suggested by [5] to evaluate the model predictive accuracy in presence of both binary and continuous response variables. In this paper, an extension of this measure to the case of discrete variables is proposed with the aim of providing a new model selection criterion when comparing different models.

The paper is organized as follows. Section 2 introduces our proposal. Section 3 illustrates the application of the proposed methodology to real data concerning cyber attacks collected at the worldwide level. Finally, the last section concludes.

2 Methodology

The proposal presented in this contribution is twofold. On the one hand, a novel model specification in the case of ordinal response variable, as is the severity variable considered in cyber risk measurement, is introduced. On the other hand, a new criterion for the comparison of different cyber risk models is illustrated.

2.1 The rank regression model

As the cyber events are typically rare and not repeatable, it is quite natural to measure them with a less demanding ordinal approach rather than using quantitative data which are often not available. Ordinal data for cyber risk measurement can be summarised, by means of a pair of statistics for each event type: the frequency of the event: how many times it has occurred, in a given period; and the corresponding severity: the mean observed loss. In the context of ordinal data, the severity can be expressed on an ordinal scale, characterised by $K = k$ distinct levels, arranged according to the corresponding magnitude. To understand the causes of cyber risks, each observed severity can be associated to a vector of explanatory variables, such as the type of attack, the technique of the attack, the victim type and the geographical area where the event has occurred.

The statistical models typically used to explain an ordinal response variable with a set of p explanatory variables are the ordered logit or probit models (see, for instance [7] and [1]). These, however, may be difficult to summarise and interpret, especially in applied contexts. We therefore develop linear regression models for a response variable that takes ordinal values. With the aim of avoiding an arbitrary assignment of the measurement scale, we resort to the ranks.

Let Y be a response variable, expressed through k ordered categories. A rank $r_1 = 1$ to the smallest ordered category of Y and a rank $(r_{j-1} + n_{j-1})$ to the following ordered categories, where n_{j-1} is the absolute frequency associated with the $(j-1)$-th category and $j = 2, \ldots, k$, are assigned. Based on this transformation, the phenomenon described by the Y variable can be re-formulated in terms of its ranks R, where:

$$R = \left\{ \underbrace{r_1, \ldots, r_1}_{n_1}, \underbrace{r_2, \ldots, r_2}_{n_2}, \ldots, \underbrace{r_k, \ldots, r_k}_{n_k} \right\}, \tag{1}$$

with $r_1 = 1$, $r_2 = r_1 + n_1$ and $r_k = r_{k-1} + n_{k-1}$.

Given p explanatory variables, a regression model for R can be specified as follows

$$\hat{r} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p, \tag{2}$$

whose unknown parameters can be estimated by the classical Maximum Likelihood method.

2.2 The *RG* as a criterion for model comparison

Typically, the specification of a model is completed by a procedure that compares different models and choose the best one in terms of goodness of fit. Here, we sug-

gest a novel metric, that takes into account the ordinal nature of the response variable. A similar measure, named *RG*, was originally provided by [5] as a criterion for the model validation in the case of both binary and continuous variables. Following [5], we extend the *RG* measure to the context of response ordinal variable.

Let R, defined in (1), be the vector of the rank-transformed response variable values and $\hat{R}$ be the vector of the corresponding predicted values. The R values can be used to build the L_R Lorenz curve (see, [6]), characterised by the following pairs: $(i/n, \sum_{j=1}^{i} r_{ord(r_j)} / \sum_{i=1}^{n} r_{ord(r_i)})$, for $i = 1, \ldots, n$, where $r_{ord(r_i)}$ indicates the rank-transformed response variable values ordered in a non-decreasing sense. Analogously, the R values can also be re-ordered in a non-increasing sense, providing the $L_R^{'}$ dual Lorenz curve.

Let $r_{ord(\hat{r}_i)}$, for $i = 1, \ldots, n$, indicate the R values re-ordered according to corresponding predicted values given by the model in (2). The set of pairs $(i/n, \sum_{j=1}^{i} r_{ord(\hat{r}_j)} / \sum_{i=1}^{n} r_{ord(r_i)})$ provides the so-called C concordance curve which measures the concordance between the response variable R and the corresponding predicted variable $\hat{R}$ orderings. In addition, the set of pairs $(i/n, i/n)$ detects the bisector curve, for $i = 1, \ldots, n$, which corresponds to the case of a random model occurring if the predicted variable values are all equal each other. For the sake of clarity, a graphical representation of the four curves is given in Figure 1.

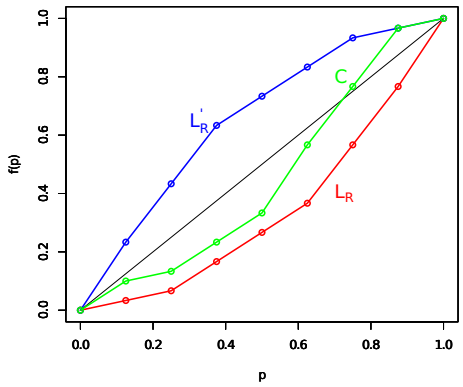

Fig. 1 The L_R (red) Lorenz curve, dual $L_R^{'}$ (blue) Lorenz curve, C (green) concordance curve and bisector curve (black).

The previous quantities can give rise to a new model selection tool, which is named *RG* as in [5]. Its formula is:

$$RG = \sum_{i=1}^{n} \frac{\left\{ (1/(n\bar{r})) \sum_{j=1}^{i} r_{ord(\hat{r}_j)} - i/n \right\}^2}{i/n}, \tag{3}$$

where $\bar{r}$ is the mean of all ranks.

A more concise expression for *RG* can also be derived as follows:

$$RG = \sum_{i=1}^{n} \frac{\left\{C(r_{ord(\hat{r}_j)}) - i/n\right\}^2}{i/n}, \tag{4}$$

where $C(r_{ord(\hat{r}_j)}) = \frac{\sum_{j=1}^{i} r_{ord(\hat{r}_j)}}{\sum_{i=1}^{n} r_{ord(r_i)}}$ is the cumulative values of the (normalised) rank-transformed response variable.

We remark that the *RG* in (3) and (4) are expressed in absolute terms. When comparing different models, a relative measure appears more appropriate making the interpretation straightforward. The relative *RG* version, denoted with RG_{norm}, can be specified as the ratio between its value and its maximum value, which is reached when the ordering of the rank-transformed response variable perfectly overlaps with the ordering of the corresponding predicted values. On the contrary, the *RG* minimum value is reached if the predicted values provided by the model are the same.

3 Application to cyber risk data

In this section we discuss an application of our proposals to cyber risk data collected by the Clusit Association, the most relevant and respected Italian association in the field of information security. The association includes, as member organisations, companies from different fields such as: Banks, Insurances, Public Administrations, Health companies and Telecommunication companies.

The data we consider consists of 6,865 worldwide observations on serious cyber attacks, in the years 2011-2017. An attack is classified as "serious" if it has led to a significant impact, in terms of economic losses and/or damages to reputation. In this paper we focus on a sample data, consisting of 808 cyber attacks observed in 2017, the year in which most data was observed. Severity levels are reported according to the type and technique of attacks (which can be seen as event types), the victims and their country of origin. We remark that the considered sample data may represent a partial situation, less critical than the real one. This because many attacks may not be disclosed, or may be disclosed very late.

Here, we focus on detecting the main factors which may affect the severity degree. For this purpose, we consider two rank regression models which differ in terms of the variables taken into account. The first rank regression model is built on all the explanatory variables appearing in our dataset. Thus, cyber attacks, attack techniques, victim type and continent are introduced into the model. A second rank regression model was specified by removing from the full model the continent variable. This in order to assess if the geographical area where the cyber attacks occur may impact on the severity degree.

In Tables 1 and 2, we report the significant effects (at a significance level $\alpha = 0.05$) provided by the full and reduced rank regression models. We remark that both models are significant yielding a *p*-value smaller than 0.001. In addition, the full

rank regression model yields $R^2 = 0.6183$ and the reduced rank regression model yields $R^2 = 0.6176$.

Table 1: Significant effects from the fitted full rank regression model at $\alpha = 0.05$.
Categorical variable reference level: cyber attack (first block): Cybercrime; victim type (second block): Automotive; attack technique (third block): 0-day

Coefficient	*Estimate*	*p-value*
Intercept	87.42	0.02678
Espionage/Sabotage	-231.38	<0.001
Hacktivism	-39.210	0.00663
Information warfare	-222.17	<0.001
Entertainment/News	117.14	0.03345
GDO/Retail	139.97	0.01743
Online Services/Cloud	136.11	0.01496
Research-Education	142.26	0.01057
Phishing/Social Engineering	120.27	0.01763
Unknown	99.670	0.04516

Table 2: Significant effects from the fitted reduced rank regression model (without continent variable) at $\alpha = 0.05$.
Categorical variable reference level: cyber attack: Cybercrime (first block); victim type (second block): Automotive; attack technique (third block): 0-day

Coefficient	*Estimate*	*p-value*
Intercept	175.65	0.01615
Espionage/Sabotage	-231.88	<0.001
Hacktivism	-38.99	0.00672
Information warfare	-221.71	<0.001
Entertainment/News	115.53	0.03549
GDO/Retail	138.18	0.01855
Online Services/Cloud	135.52	0.01514
Research-Education	140.07	0.01158
Phishing/Social Engineering	120.63	0.01708
Unknown	100.21	0.04357

From Table 1 the main interesting issue that arises from the full rank regression model is the absence of the continent variable among the significant effects. This leads us to believe that such a variable may be omitted from the model since without any impact on the cyber risk. As a further consideration, note that both models provide the same variable effect sign on the severity degree.

We now move to model validation, with the purpose of selecting the model with the highest predictive accuracy, here measured by the *RG* metric. To have a more

exhaustive picture, we also include the computation of the RMSE, as an example of traditional validation criterion. The results are displayed in Table 3.

Table 3: *RG* measure for the full and reduced (without continent variable) rank regression models

Model	*RG*	RG_{norm}	RMSE
Full rank regression model	63.185	0.739	105.196
Reduced rank regression model (without continent variable)	63.111	0.738	105.284

From Table 3, the difference between the *RG* values computed in absolute terms on the two models is really tiny. This happens also for the RMSE. In addition, since $RG_{max} \simeq 85.492$, it follows that the full and reduced rank regression models explain about the 74% of the variable ordering showing that there is no relevant difference between the models. Thus, the choice falls on the model without the continent variable.

With the aim of further validate the proposed model selection measure, we led an additional analysis in which the full rank regression model is preserved with the same variables but the reduced rank regression model is built including the variables attack techniques, victim type and continent. The only variable excluded from the model is cyber attack type. Also in this case the reduced rank regression model without the cyber attack variable is significant (p-value<0.001). The goodness of fit measure $R^2 = 0.4806$, which is greatly smaller than the R^2 value obtained on the full rank regression model. For the sake of brevity, we do not provide the table displaying the significant effects, but we only point out that, compared with the reduced rank regression model without the continent variable, in the reduced rank regression model without the cyber attack variable, Entertainment/News is no more significant while DDoS, malware, Malware and Vulnerabilities become significant.

Results in terms of *RG* and RMSE are reported in Table 4.

Table 4: *RG* measure for the full and reduced (without cyber attack variable) rank regression models.

Model	*RG*	RG_{norm}	RMSE
Full rank regression model	63.185	0.739	105.196
Reduced rank regression model (without cyber attack variable)	47.426	0.555	122.706

From Table 4, the reduced rank regression model without the cyber attack variable only explains about the 55.5% of the variable ordering showing how the role

played by the cyber attack cannot be neglected since the loss in terms of severity explanation is strongly evident. The same can be said if referring to the RMSE.

4 Conclusions

In this paper we have discussed a novel model to measure cyber risks, which takes the ordinal nature of the disclosed data correctly into account. The proposed model can be employed as a simple and effective measurement to prioritise cyber risk, as shown in our case-study. Its application to a real cyber loss database, measured at the ordinal level, reveals that the proposed tools are indeed able to detect the main factors affecting the cyber risks.

References

1. Agresti, A.: Analysis of ordinal categorical data, Second Edition, Wiley, New York (2010)
2. Clusit: 2018 Report on ICT security in Italy (2018)
3. Cox, L.A.Jr: Evaluating and improving risk formulas for allocating limited budgets to expensive risk-reduction opportunities. Risk. Anal., **32**, 1244–1252 (2012)
4. Edgar, T.W., Manz, D.O.: Research Methods for Cyber Security, Elsevier (2017)
5. Giudici, P., Raffinetti, E.: A Rank Graduation measure to assess predictive accuracy. Technical report, submitted (2019)
6. Lorenz, M.O.: Methods of Measuring the Concentration of Wealth. J. Am. Stat. Assoc. **9(70)**, 209–219 (1905)
7. McCullagh, P.: Regression Models for Ordinal Data. J. Roy. Stat. Soc. B Met. **42(2)**, 109–142 (1980)